Short Problems:

1. (This problem is inspired by Khan et al. 2011 but the specific numbers are fictitious, and probably much bigger than realistic.) We start with a standard lab strain of haploid *E. coli* which we will call "wildtype". We isolate two mutants at different loci (loci *pykF* and *topA*) and measure their fitness in all possible combinations. Assume that these two loci are the only loci that vary in our experiment. In the following table, all-caps indicates the wildtype allele.

Genotype	Relative fitness
PYKF TOPA	1.0
pykF TOPA	1.0
PYKF topA	1.1
pykF topA	1.4

   (a) If fitnesses were multiplicative, what fitness would you predict for the double mutant pykF/topA?

   (b) What kind of epistasis is this? (Positive, negative, sign?)

   (c) If we wanted to fix the advantageous *topA* mutation (via selection) as quickly as possible, would we prefer to start with $PYKF$ wildtype cells or *pykF* mutant cells? Explain briefly.

2. Consider two loci G and H in (diploid!) humans. We use pedigrees to identify haplotypes (combinations of alleles that are inherited together) in Europeans, and obtain the following data:

Haplotype	Number observed
G H	37
G h	43
g H	3
g h	17
Total	100

   (a) Calculate the disequilibrium coefficient D for these data.

   (b) What is the maximum D for these allele frequencies? (Hint: genotype frequencies must not be < 0 or > 1.)

   (c) Calculate D' = D/max(D)

   (d) The usual assumption, when we see a situation like this, is that G and H are linked on the same chromosome. Describe at least two situations in which we might see disequilibrium even if G and H were on different chromosomes, and explain each one briefly.

3. In general, what allele frequencies at two loci allow the largest possible disequilibrium D between them? Explain briefly.

4. We have a chip which can genotype 1 million common polymorphisms (SNPs) in Europeans, and use it to do genome-wide association studies (GWAS) in which we search for SNPs that are strongly correlated with a phenotype of interest, such as a disease.

   (a) A very common finding is that the SNP with the highest association with the phenotype is not the closest SNP to the mutation that actually causes the phenotype, but is a little way off. Why might this be?

   (b) Why do we prefer a European chip and not a global chip when doing GWAS in Europeans?

Long Problem:

Modern European, Asian, and Oceanian humans apparently have some ancestry from the archaic Neanderthal and Denisovan populations, whereas modern Africans do not.

The $FOXP2$ gene lies at the center of a large "archaic ancestry desert" where no modern humans appear to have Denisovan or Neanderthal alleles. Mutation in $FOXP2$ in a modern human family caused pervasive language and speech disorders.

We can suggest at least two different hypotheses for the cause of this "desert."

H1: Some time after the intermating of humans with Neanderthals and Denisovans, a favorable mutation arose in a human haplotype in this region and swept to fixation, removing any Neanderthal or Denisovan alleles that might previously have been present in the population.

H2: Prior to intermating, humans had already fixed a favorable allele in some locus in this region; thus, Neanderthal and Denisovan haplotypes were eliminated shortly after being introduced, as the human haplotypes were superior.

1. Can you suggest a way to distinguish H1 from H2, given a big sample of sequenced human genomes? In other words, what might differ in a modern human sample depending on whether H1 or H2 is correct?
2. The human and Neanderthal $FOXP2$ genes code for the same amino acid sequence (differing by two amino acids from the chimp sequence). Why doesn't this refute the idea that $FOXP2$ might be the locus responsible for the desert?
3. There are a number of deserts in the modern human genome, but this is by far the biggest. List at least two factors that could influence the size of a desert. Explain each briefly.
4. It has been suggested that modern African populations intermated with a different archaic population, but no DNA-bearing fossils of this hypothetical ancestor are available. Explain briefly why it is so much harder to be sure about archaic ancestry when the archaic sequence is unavailable. What other phenomena could be mistaken for archaic ancestry?